Faculty of Engineering and Information Technology
University of Technology Sydney

Data Mining In Epigenetic Modification and Gene Expression

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Zhixun Zhao

October 2020

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I, Zhixun Zhao declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

SIGNATURE: Production Note: Signature removed prior to publication.

[Zhixun Zhao]

DATE: 15th October, 2020

PLACE: Sydney, Australia

Acknowledgments

First and foremost, I want to thank my supervisor, Prof. Jinyan Li, for his extensive instruction and patient guidance during the last three and a half years, regardless of my research or daily life. His guidance helped me improve my research skills, such as scientific writing, academic communication, and presentation. I appreciate all his contributions of time, ideas, and funding to make all of this thesis possible.

I want to thank my advisor, Prof. Liang Fang, who advised me at my home university in China. His strong support and help gave me the chance to win the CSC (China scholarship council) scholarship and study overseas. I thank Prof. Shaoqing Li and Prof. Jihua Chen, who supervised me when I was pursuing my master's degree, for their initial guidance on my research. I thank my co-supervisor, Prof. Fang Chen, for her help in my research and providing me the opportunity to participate in industry projects.

My sincere thanks also go to my research team members: Dr. Hui peng, Dr. Yi Zheng, Dr. Yuansheng Liu, Dr. Chaowang Lan, Xiaocai Zhang, Xuan Zhang, Tao Tang, and Tian Lan, for their help in both my research and life. It's my honor to be one member of my research team, which is more like a family. Thank you all for bringing me the feeling of home and the unforgettable memories. I have special thanks to Dr. Huijun Wu, who is a truly loyal friend to me. We live together for two years and many thanks for your company.

I am also grateful to acknowledge the funding sources, including the tuition fee and living expenses provided by the China Scholarship Council

and Graduate Research School, travel funds provided by the Faculty of Engineering and Information Technologies, and vice-chancellor funding. Thanks to all staff of the Advanced Analytics Institute and School of Computer Science who provide services and conveniences to my study and research in UTS.

Lastly, I owe special thanks to my wife Zhujun Xue, my parents, and my parents-in-law. It's your love and encouragement that support me in completing my Ph.D. study. I dedicate this work to you all. Especially, I want to thank my wife for her understanding and sacrifice all these years. Love you forever.

Zhixun Zhao
October 2020 @ UTS

Contents

List of Figures

List of Tables

List of Publications

The journal and conference papers published during my PhD study are listed as follows:

Related to the Thesis :

1. **Z. Zhao**, H. Peng, J. Li et al. Imbalance learning for the prediction of N 6-Methylation sites in mRNAs [J]. BMC genomics, 2018, 19(1), 574.

2. **Z. Zhao**, H. Peng, J. Li et al. Identification of lung cancer gene markers through kernel maximum mean discrepancy and information entropy[J]. BMC Medical Genomics, 2019, 12(8): 1-10.

3. **Z. Zhao**, X. Zhang, J. Li et al. Accurate prediction of DNA N^4-methylcytosine sites via boost-learning various types of sequence features[J]. BMC genomics, 2020, 21(1), 1-11.

Others :

4. X. Zhang, **Z. Zhao**, Y. Zheng, and J Li. Prediction of Taxi Destinations Using a Novel Data Embedding Method and Ensemble Learning [J]. IEEE Transactions on Intelligent Transportation Systems, 2019.

5. H. Peng, Y. Zheng, **Z. Zhao**, J. Li et al. Recognition of CRISPR/Cas9 off-target sites through ensemble learning of uneven mis-match distributions

[J]. Bioinformatics, 2018, 34 (17), i757-i765.

6. X. Zhang, Y. Liu, Y. Zheng, **Z. Zhao**, J Li et al. Distinction Between Ships and Icebergs in SAR Images Using Ensemble Loss Trained Convolutional Neural Networks [C]. Australasian Joint Conference on Artificial Intelligence. Springer, 2018: 216-223.

7. Y. Zheng, H. Peng, X. Zhang, **Z. Zhao**, J. Li, et al. Predicting adverse drug reactions of combined medication from heterogeneous pharmacologic databases [J], BMC Bioinformatics, 2018, 19(S19).

8. Y. Zheng, H. Peng, X. Zhang, **Z. Zhao**, J. Li, et al. Old drug repositioning and new drug discovery through similarity learning from drug-target joint feature spaces [J], BMC bioinformatics, 2019, 20(23): 605.

9. Y. Zheng, H. Peng, X. Zhang, **Z. Zhao**, J. Li, et al. DDI-PULearn: a novel positive-unlabeled learning method for large-scale prediction of drug-drug interactions [J], BMC bioinformatics, 2019, 20(19): 1-12.

Abstract

This thesis employs data mining techniques to discover domain knowledge in epigenetic modification and gene expression profile. Computational methods are developed for three research questions, namely, how to accurately predict DNA N^4-methylcytosine site, how to precisely identify mRNA N^6-methyladenosine sites, and how to identify lung cancer gene expression profile markers. The motivations of the proposed methods are improving the performance of computational methods via constructing efficient feature space, optimizing machine learning schemes, solving the data imbalance issue, and employing novel statistical analysis approach to provide researchers efficient computational tools.

DNA N^4-methylcytosine (4mC) is a critical epigenetic modification and plays various roles in the restriction-modification system. The computational methods have been explored to identify 4mC in the DNA sequence in recent years due to the high cost of experimental laboratory detection. However, the state-of-the-art methods have limited performance because of the lack of effective sequence features and the ad hoc choice of learning algorithms. Chapter 3 proposes a new method with novel sequence feature space and machine learning scheme. In sequence encoding, five essential sequence features are integrated into a 292-dimension feature space, representing both global and local sequence characteristics. Then a feature selection scheme is built, where the feature importance score produced from the training process of XGBoost machine is taken as the criterion of feature selection. At last, an SVM-based prediction model is trained with the selected features

and optimized by 10-fold cross-validations. In the result part, the impact of feature selection on model performance is evaluated by an independent test. The proposed method outperforms three state-of-art predictors in both independent test and 10-fold cross-validation. Furthermore, two case studies prove the effectiveness of our method in practical situations.

N^6-methyladenosine (m^6A) widely involves in mRNA metabolism and embryogenesis. Multiple computational human mRNA m^6A site predictors have been developed. However, there are two main drawbacks of the existing methods: first, inadequate learning of the imbalanced training data; second, the sequence text features are not outstanding in representing m^6A sequence characteristics. Chapter 4 proposes to use the cost-sensitive learning idea to solve the imbalance data issues in the problem. This cost-sensitive approach learns from the entire imbalanced dataset without a random selection of negative samples. In sequence representation, site location, entropy features and specific single nucleotide polymorphism (SNP) positions are taken as new features, which improve the performs significantly. In the comparison with existing predictors, our method achieves better correctness and robustness in both independent tests and case studies. The results suggest that imbalance learning is promising to improve the performance of m^6A prediction.

The early diagnosis of lung cancer has been a challenging problem in clinical practice for a long time. The identification of differentially expressed genes as a disease marker is a promising solution. Chapter 5 presents a novel approach to identify marker genes and define the boundary of gene expression profile for human lung cancer. By calculating the kernel maximum mean discrepancy, the proposed method evaluates the expression difference between normal, normal adjacent to tumor (NAT) and tumor samples. The expression level boundaries among different groups are defined with the information entropy theory for marker genes. Compared with two conventional methods t-test and fold change, the genes selected by MMD values have better performance under all metrics in 10-fold cross-validation. Furthermore, the GO and KEGG enrichment analysis validate the discovered

marker gene in function pathways. At last, we choose ten most meaningful genes as lung cancer markers and calculate the expression profile boundaries. The proposed method is more accurate than conventional DEA methods in marker gene identification and provides a reliable method for defining the gene expression level boundaries.